Bioinformatics Advance Access published March 16, 2009

Genome analysis

Transcriptionally active gene fragments derived from potentially fast-evolving donor genes in the rice genome

Xiangfeng Wang[1,†], Zhihui Yu[2,†], Xiaozeng Yang[2], Xing-Wang Deng[1], and Lei Li[2,*]

[1]Department of Molecular, Cell and Developmental Biology, Yale University, New Haven, CT 06520, USA

[2]Department of Biology, University of Virginia, Charlottesville, VA 22904, USA

Associate Editor: Dr. Alex Bateman

ABSTRACT

The unprecedented complexity of the transcriptomic data obtained in recent years creates opportunities for new genomic studies aimed at interpolating regulatory code of gene expression and tracing genome evolution. We report here the identification and characterization of a set of 851 intergenic loci that represent transcribed gene fragments (TGFs) ectopically duplicated from 1,030 non-transposable element (non-TE) donor genes in the rice genome. We analyzed the genomic context of the TGFs and donor genes. We show that the TGFs have adopted transcriptional orientation and pattern independent of the donor genes. We further show that TGFs have undergone relaxed purifying selection, consistent with their being pseudogenized. We found that the donor genes, which are biased toward certain molecular functions, exhibit an accelerated evolution rate comparing to the genome average. Our results demonstrated a large number of actively transcribed gene fragments in the rice genome and shed light on the origin, mode of action and function of the TGFs.

1 INTRODUCTION

One of the most exciting biological findings in recent years is the discovery of widely-occurring transcriptional activity beyond the annotated genes. Transcriptome profiling efforts enabled by rapidly accumulating genome sequences and high-throughput techniques have led to the identification of numerous transcripts. One of the approaches broadly used in eukaryotic transcriptome profiling is whole genome tiling microarray that involves the construction of a virtual 'tile path' consisting of progressive oligonucleotide probes to represent a target genome (Mockler and Ecker, 2005). Tiling array analyses have been carried out in yeast (Wilhelm *et al*., 2008), fly (Stolc *et al*., 2004), worm (He *et al*., 2007), sea urchin (Samanta *et al*., 2006), human (Bertone *et al*., 2004; Cheng *et al*., 2005), Arabidopsis (Yamada *et al*., 2003), rice (Li *et al*., 2006), and legume plants (Li *et al*., 2008). Results from these studies have led inescapably to the conclusion that transcripts are produced from a large number of genomic loci that do not encode proteins or structural RNA species.

The varied properties of the newly identified novel transcripts suggest that they have diverged genomic origin and biological function. A large group of these transcripts locate in the intergenic regions and have yet to be incorporated into the molecular biology knowledge framework. One suggested mode of action for the intergenic transcripts is to provide continued transcription activity across genome regions for maintenance of an open chromatin state. Such a role is in line with results from Pol II occupancy assays where Pol II binding was found upstream of many genes that are either active or poised for rapid activation (Kim *et al*., 2005; Radonjic *et al*., 2005). In addition to influencing the structure of their surrounding chromatin, there are documented instances where intergenic transcripts act to repress transcriptional initiation at nearby genes by means of local competition between adjacent promoters (Hirschman *et al*., 1988) or interference by Pol II elongating from an upstream promoter (Martens *et al*., 2005). Thus, the function of some novel intergenic transcripts appears to require physical adjacency to their target genes.

Despite the rapid progress in our understanding of the complexity and dynamics of eukaryotic transcriptomes, relatively little attention has been paid to the interaction between physically unlinked intergenic transcripts and protein-coding genes. In the current study, we report a group of transcribed intergenic loci in the rice genome that correspond to ectopically duplicated gene fragments. Characterization of these transcribed intergenic gene fragments revealed their potential role as a mechanism contributing to functional and genomic diversity.

2 METHODS

Transcriptionally active regions (TARs) and their expression profile were obtained from a previous report (Li *et al.*, 2007). Genome information, including intergenic sequences, protein sequences, segmental duplication, GO-slim terms, and paralogous gene families were downloaded from TIGR rice genome annotation release 5 (http://rice.tigr.org/) in October 2007. The rice MULE data have been previously described (Juretic *et al*., 2005) and were downloaded from http://www.genome.org/. The Arabidopsis proteome sequences were obtained from the latest version of the Arabidopsis genome annotation (TAIR8), which we downloaded in May 2008 from www.arabidopsis.org/.

Genomic loci corresponding to intergenic TARs were translated in all six possible open reading frames and a BLASTX search was performed to establish the relationship between the deduced peptides and the annotated

*To whom correspondence should be addressed.

†These authors contributed equally to this work.

rice non-TE proteins. An intergenic locus was identified as a duplicated fragment of a donor gene if its encoded peptide (longest ORF) has an identity above 50% over a 50 amino acid span with the BLASTX expected value $E \leq 1 \times e^{-5}$. All hits were mapped against the rice genome annotation release 5 (Ouyang *et al*., 2007) again to remove loci that intersect with annotated exons or introns.

The published genomic coordinates of MULEs were based on the International Rice Genome Sequencing Project annotation version 2.0 (Juretic *et al*., 2005). We remapped all the MULEs onto TIGR annotation release 5. After mapping these sequences, TGFs that intersect with genomic regions flanked by the MULE target site duplications were identified. In addition, we also used these sequences to identify Pack-MULEs that associate with the TGFs.

In the rice genome annotation database, Plant GO-Slim Ontologies, which are a simple version of Gene Ontologies, were assigned to annotation rice proteins based on homology to Arabidopsis proteins. Hypothetical proteins, TE-related proteins and proteins with GO IDs of "unknown" definitions were excluded from this analysis. A total of 19,052 proteins have been assigned GO functions with a total of 106,119 associations made in the dataset used in the current study (Ouyang *et al*., 2007). This dataset was used as the benchmark against which the donor genes were compared. For a given GO-Slim term in the MF, CC, and BP categories, we counted the number of all rice non-TE genes and TGF donor genes linked at least once with that term. We determined whether proportions were equivalent between the two groups of genes using Pearson's chi square test with the Bonferroni's correction for multiple tests.

To carry out Ka/Ks analysis, pairs of homologous protein regions were determined for two datasets using the same criteria for identifying TGFs: (1) TGFs and donor genes in rice and (2) rice genes and Arabidopsis genes. For dataset (2), the identification cutoff was increased to 60% to reduce the number of homologous regions to facilitate downstream analysis. These pairs of homologous protein regions were aligned with CLUSTALW using default options. The alignments were subsequently superimposed on the coding region of nucleotide sequences using a Perl script based on Bioperl functions. For all pairwise alignments, the Ka and Ks substitution rate was calculated using the modified YN method as previously described (Zhang *et al*., 2006).

3 RESULTS AND DISCUSSION

Identification of TGFs

In a previous tiling microarray analysis of the *japonica* rice genome, we identified a total of 39,018 TARs including 15,472 in the intergenic regions (Li *et al*., 2007). To understand the origin and function of the intergenic TARs, we compared their deduced peptide sequences to the annotated non-TE rice protein-coding genes using a set of stringent criteria (see Section 2). This analysis identified 851 TAR-containing loci that each encodes a partial open reading frame highly similar to one or more of 1,030 annotated genes (Supplementary Table 1). These 851 partial open reading frames are referred to as putative TGFs derived from the corresponding protein-coding genes (TGF donor genes). It should be noted that the TARs are identified from a pool of four tissue types (Li *et al*., 2007) and not likely to be exhaustive. Consequently, our estimation of the number of TGFs in the rice genome is likely conservative.

To further validate TGFs and to discern their regulation, we used previously reported transcriptome data to examine the transcription pattern of TGFs and donor genes. In this dataset, a microarray containing five optimized probes for each non-TE genes and TARs in rice was used to measure gene expression in 10 different rice tissue types (Li *et al*., 2007). We detected expression of

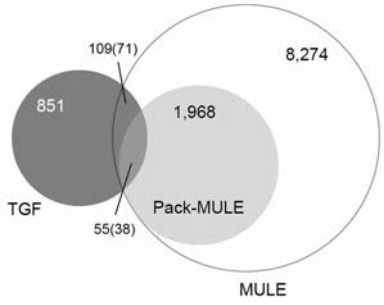

Fig. 1. Venn diagram showing the overlap between TGFs, MULEs and Pack-MULEs in the rice genome.

7,965 (18%) genes in all 10 tissue types and 28,265 (64%) genes in at least one of the 10 tissues. From the same experiment, we detected transcription of 290 (34%) and 729 (86%) TGFs in all and at least one tissue types, respectively. These results indicate that TGFs are transcriptionally as active as the protein-coding genes.

Furthermore, we found that the transcript levels of the TGFs and the corresponding donor genes are not correlated across the 10 tissue types (Pearson correlation coefficient $r = 0.20 \pm 0.13$). In contrast, transcript levels of the 2,874 genes in the 2-member paralogous gene families in rice exhibit a modest correlation ($r = 0.48 \pm 0.09$). Assuming cross-hybridization leads to elevated transcriptional correlation, our result suggests that cross hybridization between a TGF and its donor genes does not have a predominant effect on the observed transcription pattern. This conclusion in turn suggests that the TGFs are regulated independently of the donor genes.

We determined the transcriptional orientation of the TGFs relative to the ORF of the donor genes. By mapping the strandedness of all the intergenic TARs intersecting with a given TGF, we were able to determine whether a TGF is transcribed from the same strand (sense configuration), or the antisense strand (antisense configuration), or bidirectionally relative to the donor gene. From this analysis, we found that 606 (71.2%) and 414 (48.6%) TGFs are transcribed in the antisense and sense configuration, respectively, indicating antisense configuration is the dominant form of TGF transcription. A significant portion of the TGFs (169 or 19.9%) is either bidirectionally transcribed or match with more than one donor gene. Taken together, our results suggest that the TGFs have adopted transcriptional orientation and regulation independent of the donor genes. Assuming TGFs and their donor genes are transcribed in the same cell, our results suggest that they have the potential to form complex transcript networks. Furthermore, independent transcription from the TGFs and donor genes may entail a combinatorial effect on the steady state levels of the donor gene transcripts.

Chromosomal organization of TGFs

The rice genome has undergone whole genome duplications and more recently short segmental and individual gene duplications (Yu *et al*., 2005). To understand how TGFs relate to these duplication events, we examined whether TGFs are derived from tandem duplication of their donor genes. This analysis revealed that only 7 (0.7%) TGF/donor gene pairs are physically adjacent (separated by zero intervening spacer genes). In contrast, it was estimated that about 1,291 (3.1%) genes are tandem duplicates in rice using the

same criteria (Rizzon *et al.*, 2006). Thus, frequency of tandemly arrayed TGF and donor gene is significantly lower than the genome average (χ^2 test, $p < 0.001$).

We next tested whether TGFs are generated from segmental duplication by mapping TGFs and the donor genes to the known segmentally duplicated blocks in the rice genome (Ouyang *et al.*, 2007). From this analysis, we found 20 (2.0%) TGF/donor gene pairs and 2,321 (11.3%) non-TE gene pairs locate in one of the corresponding pairs of duplicated segments. Thus, TGF and donor genes are significantly depleted from segmentally duplicated blocks (χ^2 test, $p < 0.001$), indicating that the generation of TGFs occurred either independently from segmental duplications or after the duplication events.

We further examined the distribution of TGFs in the 12 rice chromosomes by comparing the number and density of TEs, non-TE genes, and TGFs for each chromosome (Supplementary Table 2). We found TGF distribution is biased toward smaller chromosomes and TGF density is strongly correlated with TE density (unit per Mb) among the 12 chromosomes (Spearman rank correlation coefficient $\rho = 0.83$, $p = 0.0058$). In contrast, TGF density and non-TE gene density is inversely correlated ($\rho = -0.89$, $p = 0.0032$).

Both Class I (retrotransposons) and II (DNA transposons) TEs are known to amplify host DNA segments. To investigate whether Class I TEs contribute to the generation of TGFs, we identified 81 TGFs whose corresponding region in the donor gene regions span introns. The majority of these TGFs (76 or 94%) retained the intron sequences. Further examination of the genomic context of the corresponding donor genes revealed that 38 (50%) of them are in fact embedded in clusters of TEs (≥ two TEs on each side) that consist mostly of Class I TEs, suggesting that TE-mediated unequal DNA exchange may have produced some of the TGFs.

Table 1. Overrepresented molecular function groups in TGF donor genes

Rank	N	X	GO-slim attribute
1	221	617	receptor activity
2	350	1103	carbohydrate binding
3	457	2753	kinase activity
4	461	2893	nucleotide binding
5	482	3283	protein binding
6	13	155	nuclease activity
7	37	786	structural molecule activity
8	72	1746	molecular function unknown
9	142	3520	hydrolase activity
10	27	2360	transcription factor activity
11	13	1740	DNA binding

Overrepresentation was determined by chi square test with Bonferroni correction for multiple tests at $p < 0.01$; N, number of donor genes; X, number of genes in the rice genome from the same GO category

Mutator-like transposable elements (MULEs) in highter plants are known to capture and carry fragments of cellular genes. Such elements are referred to as Pack-MULEs (Jiang *et al*., 2004; Juretic *et al*., 2005). In rice, whole genome scan has identified 8,274 MULEs including 1,968 Pack-MULEs. We mapped TGFs against the genomic regions flanked by the MULE target-site duplications and identified 109 (13%) TGFs that overlap with 71 MULEs, including 55 (50%) TGFs that overlap with 38 Pack-MULEs (Fig. 1). This result indicates that a significant portion of the TGFs (13%) may arise from MULE-mediated gene fragment duplication.

Functional and evolutionary implication of TGFs on donor genes

To evaluate whether the donor genes are biased toward particular functions or processes, we analyzed their assignment to the plant Gene Ontology (GO)-slim terms. We used GO-slim terms in three categories: molecular function (MF), biological process (BP), and cellular component (CC). In the predicted rice proteome, a total of 19,058 proteins have been assigned to at least one GO term with a total of 106,515 assignments (Ouyang *et al*., 2007). In the MF, BP, and CC categories, 41%, 33%, and 27% of the genes were linked to at least one GO term. In contrast, we found 784 donor genes are linked to 8,399 GO terms, with 74%, 65%, and 61% genes in the MF, BP, and CC categories, respectively (Supplementary Table 3).

Many differences were found across the three GO categories between the TGF donor genes and the whole rice proteome that were primarily caused by overrepresentation of donor genes in certain

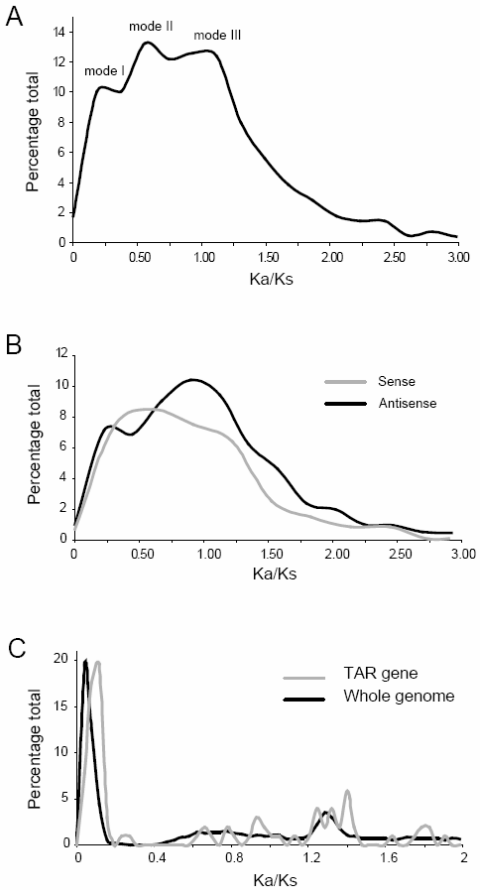

Fig. 2. Ka/Ks analysis of TGFs and donor genes. **(A)** Distribution of Ka/Ks values for pairwise comparisons of TGFs and the donor genes. (B) Distribution of Ka/Ks values for TGFs in the sense or antisense configuration against the corresponding donor genes. (C) Distribution of the Ka/Ks values for the homologous non-TE genes and homologous donor genes between rice and Arabidopsis.

GO terms (Supplementary Table 3). Significantly overrepresented GO terms in the MF category (χ^2 test, $p < 0.01$, after Bonferroni correction) are shown in Table 1. The donor genes were mostly enriched in the two types of GO terms "receptor/kinase activity" and "nucleotide/carbohydrate/protein binding" when compared with all annotated genes (Table 1). Thus, functional bias of the donor genes might have been an important determinant for the generation or retention of the TGFs.

We found 462 (34.7%) TGFs contain at least one in-frame stop codon. Thus, a large portion of TGFs has been pseudogenized. To gain further insight into the evolutionary fate of TGFs, we determined the synonymous and non-synonymous substitutions in the TGFs in comparison with the donor genes. In total, we were able to make 1,333 pairwise alignments between a TGF coding sequence and that of a donor gene. In 79 (5.9%) alignments, the TGF contained no synonymous substitution of amino acids, suggesting that these TGFs are fairly recent duplications. These alignments were not included in subsequent analysis.

We estimated the evolutionary rate of the TGFs on the basis of the ratio of non-synonymous substitutions per non-synonymous site (Ka) to synonymous substitutions per synonymous site (Ks). The average Ka/Ks value for the 1,254 TGF and donor gene pairs was 0.892 with a standard deviation of 0.702. Notably, the overall distribution of Ka/Ks values was trimodal with the three peaks locating at approximately 0.20 (mode I), 0.65 (mode II) and 1.1 (model III), respectively (Fig. 2A). When TGFs in the sense and antisense configuration were examined separately, it was found that antisense TGFs contributed mainly to mode I and III (Fig. 2B). This result suggests that while a small portion of TGFs might have been under strong purifying selection, most antisense TGFs were under neutral selection. On the other hand, the distribution of Ka/Ks for TGFs in the sense configuration contained a major peak at about 0.55 and a secondary peak around 1 (Fig. 2B). This result suggests that the sense TGFs were mainly under relaxed purifying selection and to a less extent neutral selection.

To investigate whether the presence of a TGF affects the evolution rate of the corresponding donor gene, we calculated the Ka/Ks ratio of the homologous regions between rice and Arabidopsis. When all homologous regions between the two species are considered, a bimodal Ka/Ks distribution with the major peak at 0.03 and a minor peak at 1.4 was observed (Fig. 2C). This result is consistent with the notion that while a small portion of the genes are under adaptive selection, most are under strong purifying selection. When only the homologous regions between the TGF donor genes and their Arabidopsis counterparts are considered, a bimodal distribution of Ka/Ks similar to that of the whole genome was observed (Fig. 2C). The minor peak of Ka/Ks is still at 1.4 while the major peak shifted from ~0.03 to ~0.12 (Fig. 2C). This observation indicates that although most TGF donor genes are still under strong purifying selection they potentially have a faster evolution rate than the genome average.

Given the process by which gene duplication drives genome innovation, it is conceivable that TGFs have the capability to present a foundation for the evolution of biological novelty. This novelty potential can be exploited in different ways in different lineages, dependent upon the identity of the donor genes and the TGFs. For instance, the TGFs may exert dosage effect on the donor genes. Alternatively, the TGFs may contribute to partitioning gene expression pattern of the donor genes by means of binding affinity between transcripts derived from a TGF and its homologous genes due to the accumulation of independent mutations. The transcriptional pattern of the TGFs can then impact that of its homologous genes through mechanisms such as RNA interference. Thus, regulation of gene expression by TGFs can be achieved independent of or in addition to degenerative mutations in the cis regulatory regions of the homologous genes. In particular, partitioning of expression patterns of the donor genes by TGFs provides a possible mechanism to reduce the pleiotropic constraints operating on single-copy genes by allowing synonymous mutations to tune their expression. Further functional testing of TGFs in rice and other species should provide much insight into the mechanisms by which TGFs impact donor gene function and lead to genetic diversity.

ACKNOWLEDGEMENTS

Funding: College of Arts and Sciences, University of Virginia.

REFERENCES

Bertone,P. *et al*. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306,** 2242-2246.

Cheng,J. *et al*. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308,** 1149-1154.

He,H. *et al*. (2007) Mapping the *C. elegans* noncoding transcriptome with a whole-genome tiling microarray. *Genome Res.*, **17,** 1471-1477.

Hirschman,J.E. *et al*. (1988) Genetic evidence for promoter competition in *Saccharomyces cerevisiae*. *Mol. Cell Biol.*, **8,** 4608-4615.

Juretic,N. *et al*. (2005) The evolutionary fate of MULE-mediated duplications of donor gene fragments in rice. *Genome Res.*, **15,** 1292-1297.

Kim,T.H. *et al*. (2005) A high-resolution map of active promoters in the human genome. *Nature*, **436,** 876-880.

Jiang,N. *et al*. (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, **431,** 569-573.

Li,L. *et al*. (2006) Genome-wide transcription analyses in rice using tiling microarrays. *Nat. Genet*., **38,** 124-129.

Li,L. *et al*. (2007) Global identification and characterization of transcriptionally active regions in the rice genome. *PLoS ONE*, **2,** e294.

Li,L. *et al*. (2008) Transcriptional analysis of highly syntenic regions between *Medicago truncatula* and *Glycine max* using tiling microarrays. *Genome Biol.*, **9,** R57.

Martens,J.A. *et al*. (2005) Regulation of an intergenic transcript controls adjacent gene transcription in *Saccharomyces cerevisiae*. *Genes Dev.*, **19,** 2695-2704.

Mockler, T.C. and Ecker, J.R. (2005) Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, **85,** 1-15.

Ouyang,S. *et al*. (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.*, **35,** D883-D887.

Radonjic,M. *et al*. (2005) Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon *S. cerevisiae* stationary phase exit. *Mol. Cell*, **18,** 171-183.

Rizzon,C. *et al*. (2006) Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. *PLoS Comput. Biol.*, **2,** e115.

Samanta,M.P. *et al*. (2006) The transcriptome of the sea urchin embryo. *Science*, **314,** 960-962.

Stolc,V. *et al*. (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science*, **306,** 655-660.

Wilhelm,B.T. *et al*. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453,** 1239-1243.

Yamada,K. *et al*. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science*, **302,** 842-846.

Yu,J. *et al*. (2005) The genomes of *Oryza sativa*: A history of duplications. *PLoS Biol.*, **3,** e38.

Zhang,Z. *et al*. (2006) Computing Ka and Ks with a consideration of unequal transitional substitutions. *BMC Evol. Biol.*, **6,** 44.